

# *Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae*

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Shaw, L. P. ORCID: <https://orcid.org/0000-0001-7332-0820>,  
Chau, K. K. ORCID: <https://orcid.org/0000-0002-6926-080X>,  
Kavanagh, J. ORCID: <https://orcid.org/0000-0003-1495-0843>,  
AbuOun, M. ORCID: <https://orcid.org/0000-0003-4190-1473>,  
Stubberfield, E. ORCID: <https://orcid.org/0000-0002-5527-1209>, Gweon, H. S. ORCID: <https://orcid.org/0000-0002-6218-6301>, Barker, L., Rodger, G. ORCID: <https://orcid.org/0000-0001-6051-8985>, Bowes, M. J. ORCID: <https://orcid.org/0000-0002-0673-1934>, Hubbard, A. T. M., Pickford, H., Swann, J. ORCID: <https://orcid.org/0000-0002-1475-8248>, Gilson, D. ORCID: <https://orcid.org/0000-0002-3713-1966>, Smith, R. P., Hoosdally, S. J., Sebra, R., Brett, H. ORCID: <https://orcid.org/0000-0002-4100-8565>, Peto, T. E. A., Bailey, M. J., Crook, D. W. ORCID: <https://orcid.org/0000-0002-0590-2850>, Read, D. S. ORCID: <https://orcid.org/0000-0001-8546-5154>, Anjum, M. F., Walker, A. S. ORCID: <https://orcid.org/0000-0002-0412-8509> and Stoesser, N. ORCID: <https://orcid.org/0000-0002-4508-7969> (2021) Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. Science Advances, 7 (15). eabe3868. ISSN 2375-2548 doi: <https://doi.org/10.1126/sciadv.abe3868> Available at

<https://centaur.reading.ac.uk/97619/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1126/sciadv.abe3868>

Publisher: AAAS

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

[www.reading.ac.uk/centaur](http://www.reading.ac.uk/centaur)

## **CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

## GENETICS

# Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae

Liam P. Shaw<sup>1\*</sup>, Kevin K. Chau<sup>1</sup>, James Kavanagh<sup>1</sup>, Manal AbuOun<sup>2</sup>, Emma Stubberfield<sup>2</sup>, H. Soon Gweon<sup>3,4</sup>, Leanne Barker<sup>1,5</sup>, Gillian Rodger<sup>1,5</sup>, Mike J. Bowes<sup>3</sup>, Alasdair T. M. Hubbard<sup>1,6</sup>, Hayleah Pickford<sup>1,5</sup>, Jeremy Swann<sup>1,7</sup>, Daniel Gilson<sup>8</sup>, Richard P. Smith<sup>8</sup>, Sarah J. Hoosdally<sup>1</sup>, Robert Sebra<sup>9,10</sup>, Howard Brett<sup>11</sup>, Tim E. A. Peto<sup>1,5,7</sup>, Mark J. Bailey<sup>3</sup>, Derrick W. Crook<sup>1,5,7</sup>, Daniel S. Read<sup>3</sup>, Muna F. Anjum<sup>2</sup>, A. Sarah Walker<sup>1,5,7</sup>, Nicole Stoesser<sup>1,5\*</sup>, REHAB consortium

Copyright © 2021  
The Authors, some  
rights reserved;  
exclusive licensee  
American Association  
for the Advancement  
of Science. No claim to  
original U.S. Government  
Works. Distributed  
under a Creative  
Commons Attribution  
License 4.0 (CC BY).

*Escherichia coli* and other Enterobacteriaceae are diverse species with “open” pangenomes, where genes move intra- and interspecies via horizontal gene transfer. However, most analyses focus on clinical isolates. The pangenome dynamics of natural populations remain understudied, despite their suggested role as reservoirs for antimicrobial resistance (AMR) genes. Here, we analyze near-complete genomes for 827 Enterobacteriaceae (553 *Escherichia* and 274 non-*Escherichia* spp.) with 2292 circularized plasmids in total, collected from 19 locations (livestock farms and wastewater treatment works in the United Kingdom) within a 30-km radius at three time points over a year. We find different dynamics for chromosomal and plasmid-borne genes. Plasmids have a higher burden of AMR genes and insertion sequences, and AMR-gene-carrying plasmids show evidence of being under stronger selective pressure. Environmental niche and local geography both play a role in shaping plasmid dynamics. Our results highlight the importance of local strategies for controlling the spread of AMR.

## INTRODUCTION

Enterobacteriaceae are a family of Gram-negative bacteria that can cause clinical infections (1, 2) and persist environmentally (3, 4) across diverse environmental niches (5). Antimicrobial resistance (AMR) in Enterobacteriaceae has emerged as a major problem in the past decade (6, 7). Dissemination of AMR genes often occurs via mobile genetic elements (MGEs), which can transfer genes within and between species both locally (8) and globally (9). Freshwater-, wastewater-, and livestock-associated strains of Enterobacteriaceae have been proposed as reservoirs for AMR genes in clinical isolates (10–13), but the links between these remain cryptic (14).

Species within Enterobacteriaceae are well-known examples of “open” pangenomes (15, 16) containing substantial genetic diversity, with movement of genes via horizontal gene transfer (HGT) (17). Current understanding of the ecology and evolution of pangenomes is incomplete (18), with ongoing debate about the roles of niche adaptation and selection (19–22). However, published Enterobacteriaceae genomes are biased toward clinical isolates (23, 24), and

sampling frames reflecting truly interlinked communities are limited. Much remains unknown about the population genetics of Enterobacteriaceae (25) and the role of plasmids in nonclinical contexts (26).

Genomic studies of Enterobacteriaceae have predominantly used short-read whole-genome sequencing. AMR genes and their flanking regions are frequently fragmented in short-read assemblies due to repetitive elements and structural rearrangements (8, 27). Combining short and long reads (“hybrid assembly”) produces complete, high-quality genomes (28), allowing accurate structural resolution. Here, we report a study of the multispecies pangenome within nonclinical Enterobacteriaceae using hybrid sequencing. We used hybrid assembly of 827 sympatric Enterobacteriaceae (including *Citrobacter*, *Enterobacter*, *Escherichia*, and *Klebsiella* spp.) to characterize the pangenome of these genera, considering both niche [cattle, pig, sheep, or wastewater treatment work (WwTW) associated] and geography (sampling location).

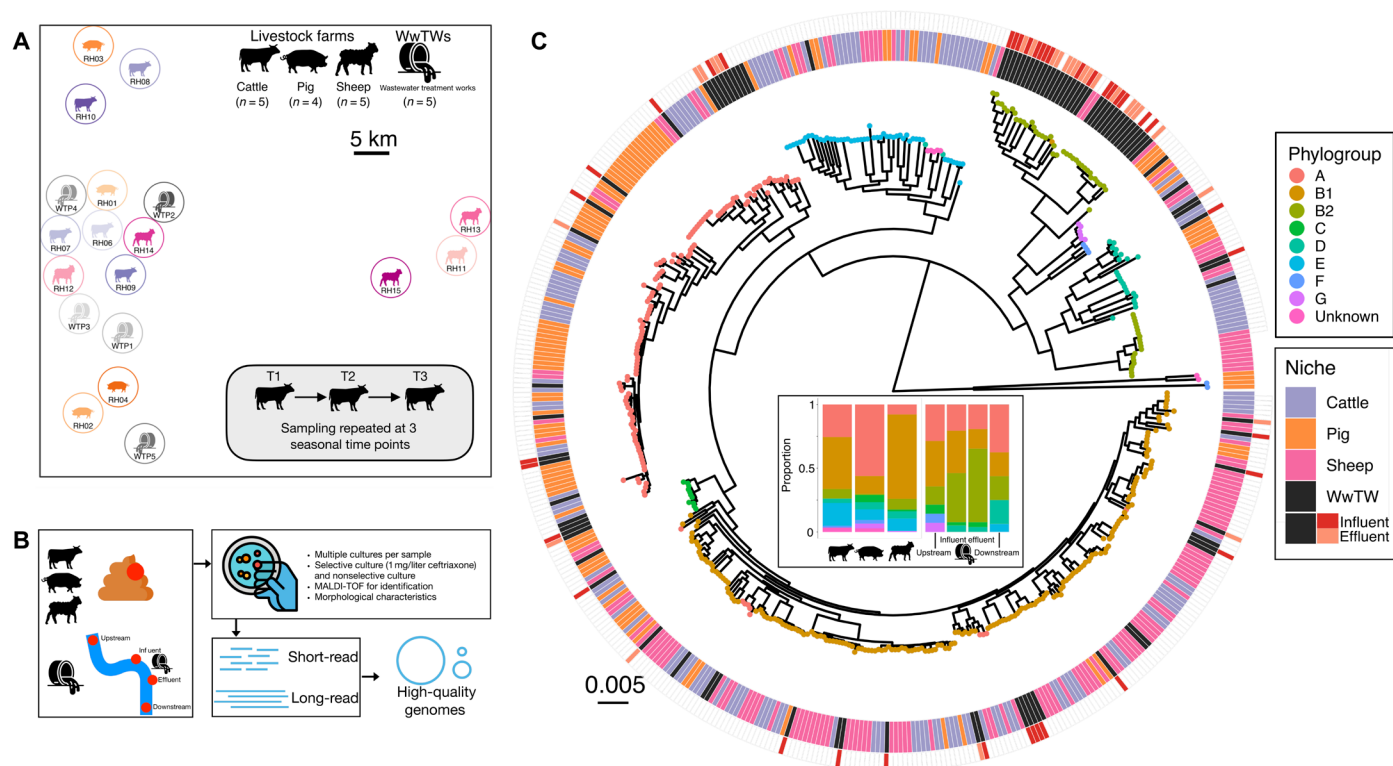
## RESULTS

### A diverse collection of complete genomes from livestock and water-borne niches

We collected samples from 19 locations  $\geq 5$  km apart (maximum distance: 60 km) in South Central England (United Kingdom) in 2017, namely, 14 livestock farms (four pig, five cattle, and five sheep) and water sources around five WwTWs over three seasonal time points (TPs) (Fig. 1A). We pooled samples within each location and TP and then cultured isolates from these pooled samples. A selected subset of 832 of 2098 cultured isolates underwent short- and long-read sequencing and hybrid genome assembly (Fig. 1B; see Materials and Methods), resulting in 827 high-quality genomes (table S1;  $n = 495$  from livestock farms,  $n = 332$  from WwTWs), from four genera: *Citrobacter* ( $n = 127$ ), *Enterobacter* ( $n = 71$ ),

<sup>1</sup>Nuffield Department of Medicine, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK. <sup>2</sup>Department of Bacteriology, Animal and Plant Health Agency (APHA), Woodham Lane, Addlestone, Surrey KT15 3NB, UK. <sup>3</sup>UK Centre for Ecology & Hydrology (UKCEH), Benson Lane, Crowmarsh Gifford, Wallingford OX10 8BB, UK. <sup>4</sup>School of Biological Sciences, University of Reading, Reading RG6 6AS, UK. <sup>5</sup>NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK. <sup>6</sup>Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK. <sup>7</sup>NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford OX4 9DU, UK. <sup>8</sup>Department of Epidemiological Sciences, The Animal and Plant Health Agency (APHA), Woodham Lane, Addlestone, Surrey KT15 3NB, UK. <sup>9</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>10</sup>Sema4, a Mount Sinai venture, 333 Ludlow Street, North Tower, 8th floor, Stamford, CT 06902, USA. <sup>11</sup>Thames Water Utilities, Clearwater Court, Vastern Road, Reading RG1 8DB, UK.

\*Corresponding author. Email: liam.philip.shaw@gmail.com (L.P.S.); nicole.stoesser@ndm.ox.ac.uk (N.S.)



**Fig. 1. Overview of the diverse *Escherichia coli* isolates in this study.** (A) Relative sampling locations of the farms (cattle, pig, and sheep) and wastewater treatment plants (WwTWs) in this study, sampled at three different TPs. (B) Schematic illustration of the sampling, culture, and sequencing workflow, resulting in high-quality genome assemblies with a median of one circularized chromosome and two circularized plasmids per assembly. (C) Mid-point rooted core genome phylogeny of *E. coli* isolates ( $n = 488$ ), with tips colored by phylogroup and ring colors showing sampling niche. Inset panel at center of phylogeny shows phylogroup abundances (as proportion of isolates) from different sampling niches.

*Escherichia* ( $n = 553$ ), and *Klebsiella* ( $n = 76$ ). Most farm isolates were *Escherichia* spp. (451 of 495, 91.1%), with WwTW isolates having roughly even proportions of genera (fig. S1). Isolates contained a median of one AMR gene (range: 0 to 23), with variation by genus: *Klebsiella* isolates carried a median of 4 (range: 1 to 18).

Isolates were highly diverse, containing diversity not present in published genomes (fig. S2). *Escherichia* diversity included all main *E. coli* phylogroups (Fig. 1C), as well as 53 *Escherichia fergusonii*, and 13 isolates from clades I, II, III, and V (table S1). Phylogroup B2 was strongly associated with WwTWs compared with livestock (34.3% versus 5.3% of *Escherichia* isolates in each category respectively;  $\chi^2 = 70.4$ ,  $P < 0.001$ ), particularly in influent and effluent samples (Fig. 1C). This observation is in line with Ludden *et al.* (29), who found that phylogroup B2 was rare in livestock-associated *E. coli* sampled in England (4 of 431 genomes). Pigs had a greater proportion of phylogroup A isolates (Fig. 1C). Of 187 identified *E. coli* multilocus sequence types (STs), 56.1% (105 of 187) were seen only once, similar to the 61% observed by Touchon *et al.* (30) in a study of nonclinical *E. coli*. Only 12 *Escherichia* STs were seen in both livestock and WwTW isolates, with phylogroup B1 as the most represented (5 of 12 STs). ST10 was the most prevalent ST ( $n = 45$ ), seen in 10 of 14 farms and 3 of 5 WwTWs. This is in agreement with recent studies in England: Ludden *et al.* (29) also found ST10 to be the most prevalent *E. coli* ST in livestock-associated isolates, and AbuOun *et al.* (13) found that the ST10 clonal complex was the most prevalent in isolates from pig farms. Our observations

and good agreement with recent work in this setting suggest that our dataset, although sampled from a relatively small geographical region, is representative of broader nonclinical *E. coli* populations.

Considering only livestock *E. coli* isolates, over time, there was a persistent phylogroup signature of both livestock host and farm, with individual farm explaining slightly more variance than livestock type ( $R^2 = 28.1\%$  versus  $25.8\%$ ; fig. S3). In other words, between-farm differences in *E. coli* phylogroup abundance were of the same order as differences between livestock species. However, livestock type explained less variance for STs than phylogroups ( $R^2 = 8.5\%$ ), with only 39 of 131 STs (29.8%) seen on more than one farm. This suggests that associations between livestock species and *E. coli* population structure are more important at the higher level of phylogroup, and clear associations between ST and livestock species are not the norm. There were only 26 instances where an *E. coli* ST was observed over time on the same farm (involving 16 STs), and most of these (22 of 26 instances, 12 of 16 STs) were STs also seen across farms (fig. S4); these could therefore represent more generalist STs.

However, STs can encompass large genomic variation. We therefore also considered *E. coli* strain clusters using a core genome distance of <100 single-nucleotide variants (SNVs) (maximal diversity observed across sampled *E. coli*: 211,251 SNVs; median pairwise distance: 46,144 SNVs). There were 280 isolate pairs with <100 SNVs, of which 181 (64.6%) were isolates cultured from the same pooled sample (i.e., same farm and same TP) (fig. S5A). Overall, 10.5% of all isolate pairs from the same pooled sample had <100

SNVs between them, compared with 1.4% ( $n = 52$ ) of isolate pairs from different TPs on the same farm and 0.2% ( $n = 44$ ) between different farms of the same animal (fig. S5B). Of the latter, 41 of 44 were between cattle farms, and 36 involved a single cattle farm (RH06). There were only three isolate pairs with <100 SNVs between farms of different animals (fig. S5A). All of these were between farms in close geographic proximity (two instances from pig farm RH03 and cattle farm RH10, and one instance from cattle farm RH07 and sheep farm RH12; see Fig. 1A for distances), suggesting local strain movement. There were no isolate pairs with <100 SNVs between WwTW and livestock niches, and only three isolate pairs occurred across TPs at WwTWs (all at a single WwTW).

Together, these results indicate that different livestock hosts have a stable balance of *E. coli* phylogroups and that each farm setting can harbor substantial strain-level diversity, which, in our dataset, is unique to that farm and can persist over time. In contrast, isolates from locations proximal to WwTWs do not have a stable population of strains, reflecting the more transitory nature of this setting. We did not recover any transmission links at the strain level between WwTWs and livestock, although the vast diversity of natural *E. coli* populations means this should not be interpreted as evidence that these links do not exist.

### Plasmid gene repertoires are linked to genus and niche

We recovered 2292 circularized plasmids across all Enterobacteriaceae, ranging in size from 1240 bp to 824 kbp (median: 43 kbp; table S2). There were 297 of 2292 (13.0%) with no identifiable plasmid replicon, and most of these were from WwTW isolates (192 of 297, 64.6%). Multiple replicons were carried by 723 of 2292 (31.5%), and these plasmids tended to be larger (median length: 106,811 bp versus 6669 bp for single replicon plasmids). Of *E. coli* isolates with complete genomes, over two thirds (70.4%, 245 of 348) carried a plasmid with an IncFII replicon. Forty-three percent of circularized plasmids (986 of 2293) had at least one match with >99% identity to other publicly available plasmid sequences (fig. S2B). However, 12.3% (282 of 2293) had a top identity score of <95% to a previous known sequence (fig. S2B), and 17 plasmids with no match were identified, suggesting that sampling recovered previously undescribed plasmid diversity. We grouped circularized plasmids into 611 distinct plasmid clusters using alignment-free distances (see Materials and Methods), which closely matched their gene content (fig. S6A). A recent analysis by Redondo-Salvo *et al.* (31), clustering over 10,000 plasmids from prokaryotes using average nucleotide identity, found that plasmids within the same cluster contained a common genomic backbone. We also found that the synteny of shared genes was strongly conserved within plasmid clusters (fig. S6B), supporting this concept of common backbones that remain stable while allowing genes to be gained by insertion.

A median of 3.3% of genes were on plasmids (range: 0 to 16.5%), with substantial variation by genus and niche (fig. S7A). This is a comparable value to the 3% finding of Touchon *et al.* (30) in *E. coli* with variation over a similar range [Fig. 4A of (30)]. We also considered the effect of plasmid copy number (i.e., multiplying plasmid lengths by their inferred copy number) to calculate the total DNA in both chromosomes and plasmids within isolates (fig. S7B). *E. coli* isolates had a median of 5.7% of DNA present on plasmids, which was substantially higher in pig farm isolates (median: 10.1%; fig. S7B), linked to the presence of large plasmids. We analyzed the variation of gene content with phylogeny and niche (Fig. 2). Chromosomal

genes were highly genus specific ( $R^2 = 55.0\%$ ); the plasmid-borne pangenome was far more variable but still had a weak association with genus ( $R^2 = 6.5\%$ ) (Fig. 2). Within *E. coli*, plasmid gene content was linked to niche ( $R^2 = 5.6\%$ ) and phylogroup ( $R^2 = 5.2\%$ ), with a stronger interaction between niche and phylogroup ( $R^2 = 7.9\%$ ) (Fig. 2).

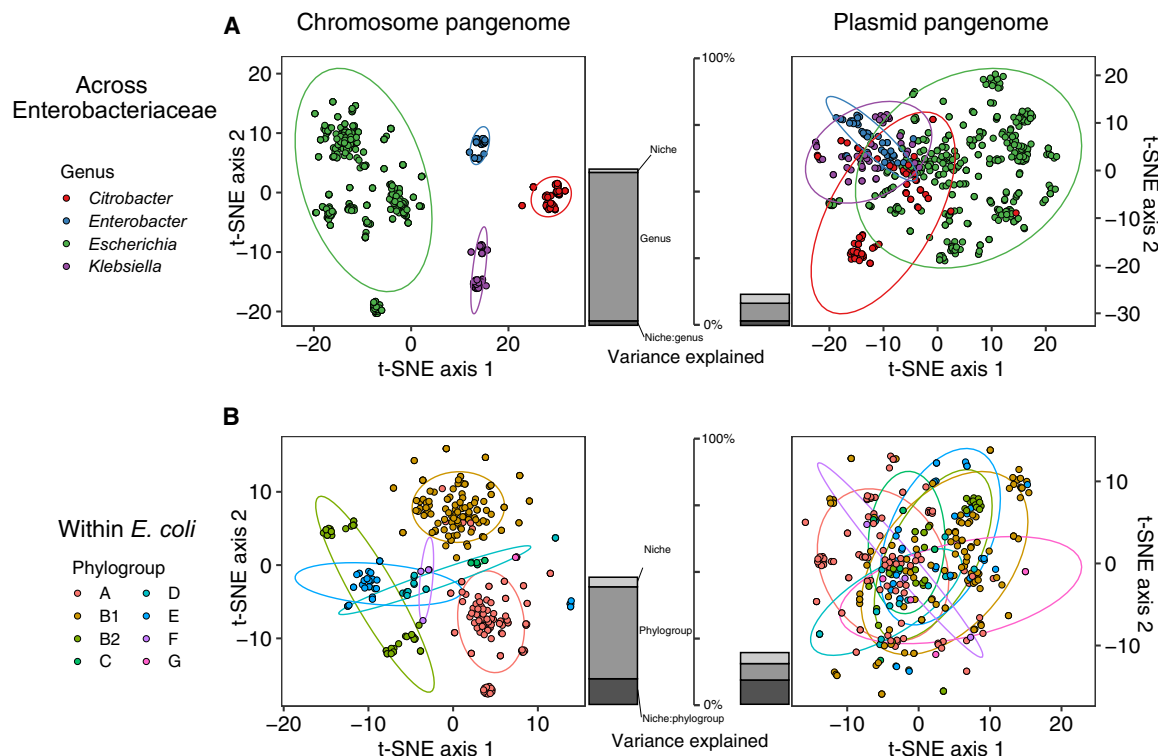
Plasmids were predicted to be conjugative, mobilizable, or non-mobilizable (see Materials and Methods), and we explored whether this affected their distribution. Nonmobilizable plasmid clusters were less commonly shared between different phylogroups within farms compared with mobilizable or conjugative plasmids (fig. S8). Although AMR genes were predominantly found in conjugative or mobilizable plasmid clusters, consistent with their frequent acquisition and movement between strains, plasmid clusters with AMR genes were not more commonly distributed across multiple phylogroups (chi-square test  $\chi^2 = 0.64$ ,  $P = 0.42$ ; fig. S8). However, on pig farms, most of the conjugative plasmid clusters seen across multiple phylogroups carried AMR genes, suggesting that they play an important role within this niche.

Since many isolates had multiple plasmids in their genome, we also considered the possibility of plasmid-plasmid interactions. Positive epistasis between large (>10 kbp) and small plasmids has been suggested to promote plasmid stability in Enterobacteriaceae, based on analyses of genomes in public databases (32). In *E. coli* isolates with complete genomes ( $n = 348$ ), we observed a significant association between small and large plasmid presence (chi-square test  $\chi^2 = 4.44$ ,  $P = 0.035$ ), with 45.7% carrying at least one large (>10 kbp) and one small plasmid and only 3.7% carrying a small plasmid without a large plasmid. Thus, previous results that support the existence of positive large-small plasmid epistasis are reproducible in this natural population, suggesting that this is an important feature of Enterobacteriaceae plasmid distributions. We also found evidence for specific plasmid-plasmid associations. For example, cattle *E. coli* isolates showed co-occurrence of a ColRNA plasmid (cluster 37: median length, 4.6 kbp) and an IncFII plasmid cluster (cluster 279: median length, 106 kbp), with 14 of 16 isolates with the ColRNA plasmid also carrying the larger IncFII plasmid. Isolates were from three phylogroups (A:  $n = 2$ , B1:  $n = 5$ , and E:  $n = 9$ ) and four farms, suggesting a robust association that reflects plasmid epistasis independent of chromosomal background.

### Plasmids carry an overrepresentation of AMR genes and insertion sequences

Plasmids carried more diverse and less genus-restricted genes, as expected from their role as the more flexible component of the genome. Despite carrying just 3.3% of total gene content, plasmid-borne genes accounted for 11.5% of unique genes (8.9 to 17.0% considering each genus; fig. S9), and 40.1% were seen in more than one genus (19.6 to 55.6% considering each genus; table S3). Plasmids are important vehicles of AMR genes in clinical contexts; similarly, plasmids had a much greater burden of AMR genes in the niches studied here. Considering isolates with circularized chromosomes (see Materials and Methods), 901 of 1876 AMR genes (48.0%) were found on plasmids, i.e., a 14.5× relative burden in plasmids. They also had a higher burden of insertion sequences (ISs), which are linked to the movement of genes by HGT. Of 26,565 ISs, 3695 (21.7%) were found on plasmids (6.6× relative burden). There was a weak correlation between the number of plasmid- and chromosome-associated AMR genes within an isolate (Spearman's  $\rho = 0.11$ ,





**Fig. 2. The plasmid-borne component of the pangenome is structured by niche and phylogeny, with greater variation than in the chromosomal component.**

Plots are shown for isolates (A) across Enterobacteriaceae and (B) within *E. coli*, for both the chromosomal component of the pangenome and the plasmid-borne component analyzed separately. Color indicates (A) genus within Enterobacteriaceae and (B) phylogroup within *E. coli*. Stacked bar charts in the center of each show the variance in gene content explained by niche, phylogeny (genus or phylogroup), and their interaction. The plasmid-borne component has greater residual variance than the chromosomal component, with a comparatively stronger niche-phylogeny interaction (darkest shaded bar).

$P = 0.004$ ) but a strong positive correlation for the number of ISs (Spearman's  $\rho = 0.40$ ,  $P < 0.001$ ) (fig. S10A) seen across genera (fig. S10B).

We observed different patterns of ISs across chromosomes and plasmids (fig. S11). Some ISs were strongly associated with plasmids, with the strongest association being for IS26. However, 27.5% of isolates carrying IS26 on a plasmid also carried it on their chromosome, consistent with its characteristically active behavior. It has been shown that IS26 can drive the reorganization of plasmids in clinical settings by replicative transposition (33), as well as amplify AMR genes (34) and create within-plasmid heterogeneity in a single isolate (35). Its high prevalence in Enterobacteriaceae plasmids here suggests that it may play a similar role outside the context of clinical AMR. The most prevalent IS on both chromosomes and plasmids was ISKpn26, with 50.2% of ISKpn26-positive isolates having it both chromosomally and plasmid borne.

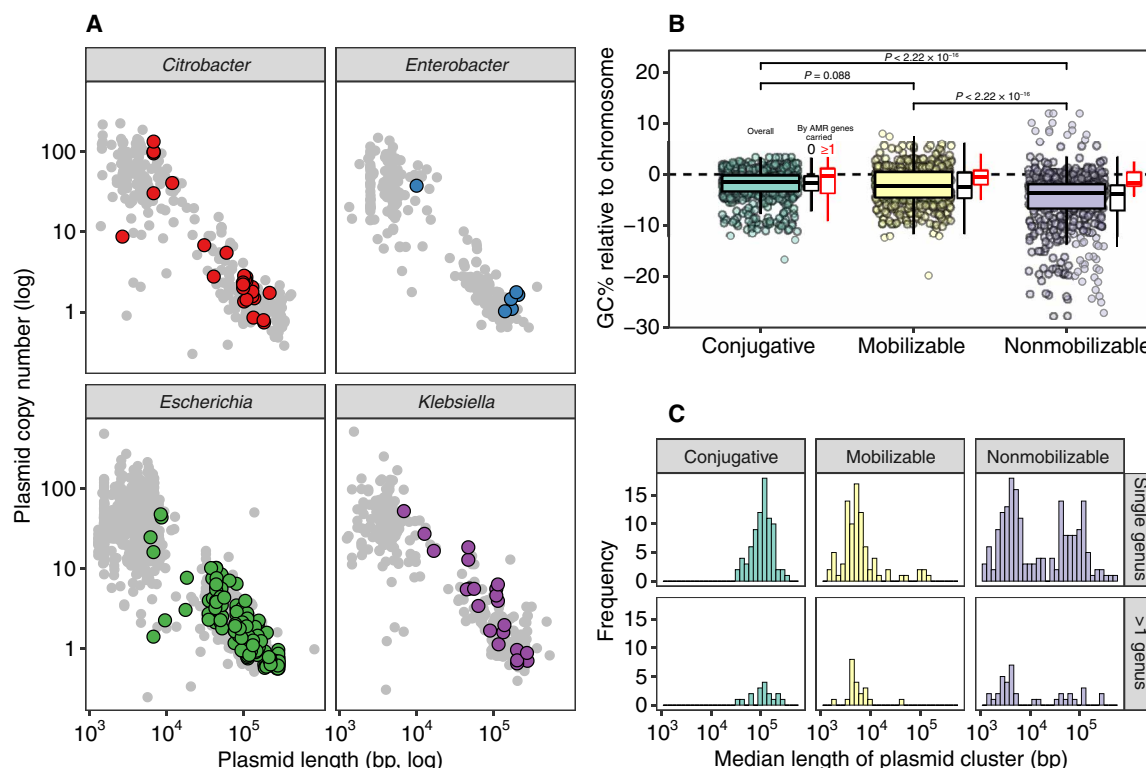
Considering *Escherichia*, WwTW isolates showed a greater diversity of ISs, with 65% of ISs found in a higher proportion of WwTW isolates compared with those from farms (fig. S12), including IS30 which has been proposed as a marker for naturalized wastewater populations of *E. coli* (36). Touchon *et al.* (30) suggest that water-borne *E. coli* strains are adapted to this niche rather than being fecal contaminants and can therefore adapt by acquiring genetic material from not only gut bacteria (as in, e.g., a livestock or human host) but also other diverse environmental bacteria. The observation of greater IS diversity in WwTW isolates here would be consistent with this hypothesis.

We also investigated the overall patterns of co-occurrence of ISs to see whether they were strongly linked on plasmids. Overall, ISs had random levels of co-occurrence on *Escherichia* plasmids (upper tail  $P = 0.85$  from null model simulations of checkerboard score; see Materials and Methods; fig. S13A), suggesting that ISs frequently move independently between plasmid backgrounds. In contrast, applying the same method to AMR genes, we found they significantly co-occurred (upper tail  $P = 0.02$ ; fig. S13B), suggesting coselection and underlining the tendency of some AMR genes to co-occur in specific regions of plasmids.

### Plasmids carrying AMR genes show features suggestive of selection

Plasmids fell into two broad classes across genera: small multicopy plasmids (<10 kbp, 10× to 100× copy number inferred from coverage relative to chromosome) and large low-copy plasmids (>10 kbp, <10×) (Fig. 3A). AMR plasmids were almost all large low-copy plasmids (172 of 183, 94.0%). While small multicopy plasmids are of interest in facilitating evolutionary innovation (37, 38), this finding suggests that they do not play a major, direct role in AMR in livestock- and WwTW-associated Enterobacteriaceae.

Experimental evidence shows that selective advantages favor lower GC content in intracellular elements (39), and this has been proposed as an explanation for the tendency of plasmids to typically have lower GC content than their host chromosome. Under this hypothesis, the more dependent on the bacterial host the plasmid is, the lower its relative GC content should be. We investigated



**Fig. 3. Distinct plasmid lifestyles between AMR and non-AMR plasmids.** (A) Plasmid length (x axis) and inferred copy number (y axis) of all circularized plasmids ( $n = 2292$ ), faceted by genus. Plasmids with  $\geq 1$  AMR gene (colored points) tended to be larger and present in lower copy numbers. (B) Relative GC content of all plasmids to their host chromosome for all circularized plasmids present in an assembly with a circularized chromosome ( $n = 1753$  plasmids across 616 isolates), split by predicted plasmid mobility. Boxplots are additionally shown classifying plasmids within predicted mobility types by the number of AMR genes carried: those  $\geq 1$  AMR gene (red) or no AMR genes (black). Comparisons with  $P$  values are shown for all plasmids within a predicted mobility class. (C) Length distributions of plasmid clusters (see Materials and Methods).

this in our dataset. Overall, plasmids had a lower relative GC content than their host chromosomes (median difference, 2.5%; Fig. 3B). There was an approximate gradient of relative GC content with predicted plasmid mobility (Fig. 3B), with plasmids predicted to be mobile having a smaller relative difference. Furthermore, this difference was less marked for AMR plasmids (median, 0.3%) across mobility categories (Fig. 3B). Nearly half had a higher GC content than their host chromosome (46.7% versus 17.7% of non-AMR plasmids). Together, this suggests that AMR plasmids are being selected for in these environments, which counteracts the usual selective advantages for lower GC content. Alternatively, this may be a signal of their relatively recent acquisition by their host.

### Evidence for recent HGT across genera and within isolates

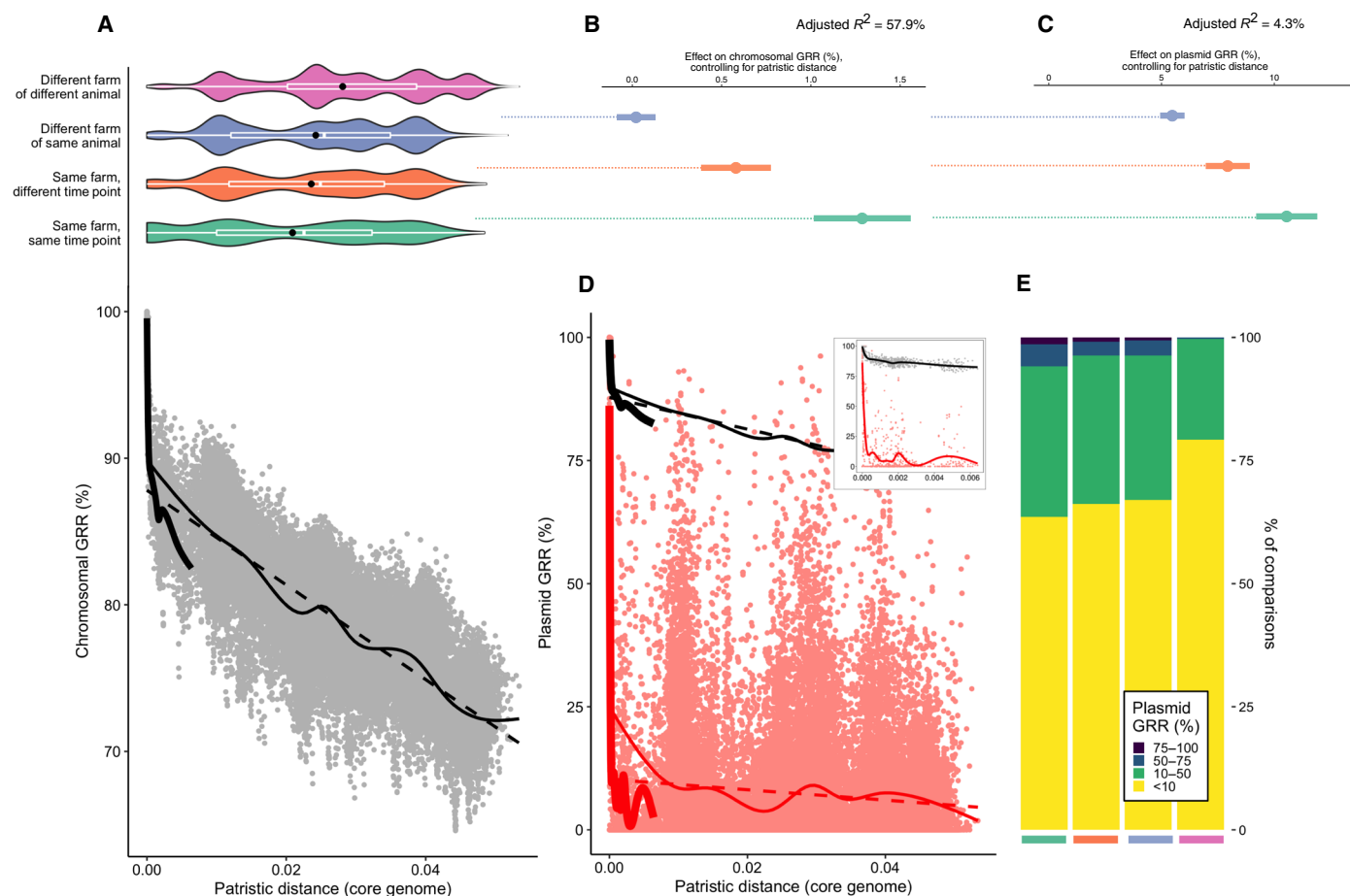
We identified 2364 potential HGT events involving transfers of sequence  $>5000$  base pairs (bp) between isolates of different genera (see Materials and Methods). These represent possible instances of the recent movement of genetic material across species boundaries. Isolates from the same farm were  $\sim 10\times$  more likely to show evidence of cross-genera HGT than would be expected (chi-square test  $\chi^2 = 1159$ ,  $P < 0.001$ ; fig. S14), and 12.3% of these cross-genera HGT events involved at least one AMR gene, with most of these AMR HGT events between pig isolates (37 of 48, 77.0%).

The movement of genes can also occur within individual genomes. We therefore also investigated occurrences where the same

gene was present on both the chromosome and plasmid(s) within an *E. coli* genome. We observed distinct differences between niches, with increased amounts of chromosome-plasmid sharing in pig and WwTW isolates compared with cattle and sheep (fig. S15). This may be a signature of increased selection for AMR in these niches, such that usually transitory gene movements and duplications are retained in genomes for long enough to be detected.

### Quantifying the roles of phylogeny, niche, and geography in the *E. coli* pangenome

To understand the strength of different factors shaping the pangenome, we analyzed the pangenome of a single species, *E. coli*, in more detail. Isolates recovered from the same location spanned total *E. coli* diversity (Fig. 4A). Interisolate core genome distances were strongly correlated with chromosomal gene repertoire relatedness (GRR) (Fig. 4A). Core genome distance explained the majority of variance in chromosomal GRR (Fig. 4B), but there was a consistent contribution from geography and time: isolates from the same pooled sample sharing more genes than would be expected (+1.2%), as did isolates from the same farm at different time points (+0.5%) (Fig. 4B). There was no such effect for isolates from different farms of the same livestock, suggesting that this reflects local geography rather than adaptation to livestock host. Although the variance explained was much lower, local geography effects were also observed for plasmid GRR (Fig. 4C), but core genome distance was



**Fig. 4. The interplay of phylogeny and niche in the *E. coli* pangenome.** (A) Pairwise comparisons of GRR for chromosomal genes show that chromosomal GRR falls off rapidly at small patristic distances, followed by an approximately linear decrease. Fits show intra-ST comparisons (thick black line), all comparisons (thin black line), and a linear model (dashed black line). Violin plots above show the distribution of patristic distances depending on the relative sample source of the two isolates in the pairwise comparison (white boxplot: median and IQR; black point: mean), showing that even isolates cultured from the same sample (same farm and same TP) span equivalent diversity to isolates cultured from different locations. (B) Coefficients from a linear model for chromosomal GRR with an interaction term with patristic distance (excluding intra-ST comparisons). (C) Variance explained by phylogeny and geography for chromosomal and plasmid GRR. (D) GRR for plasmid-borne genes with patristic distance. Fits show intra-ST comparisons (thick red line), all comparisons (thin red line), and a linear model (dashed red line). Inset panel shows left-hand region of the plot with only intra-ST comparisons, with chromosomal GRR relationship also shown (gray points, black line). (E) Plasmid GRR comparisons shown by isolate sources, excluding intra-ST comparisons. Colors on the x axis are the same as in (A). Plots include all *E. coli* isolates with a circularized chromosome ( $n = 363$ ).

uncorrelated with plasmid GRR apart from for near-identical strains (Fig. 4D). Isolates from different STs from different farms of the same livestock could still have high plasmid GRR (Fig. 4E), suggesting that host-specific plasmids may facilitate niche adaptation.

## DISCUSSION

We have investigated the pangenome of major genera of sympatric Enterobacteriaceae from locations within a 30-km radius, using a diverse set of nonclinical isolates cultured from the same samples and focusing in detail on *E. coli*. Despite high overall diversity, with most of the strains only observed once in the dataset, we observed the persistence of strains and plasmids on farms over the course of a year. Our results highlight the combination of persistence and dynamism that characterizes Enterobacteriaceae genomes at multiple scales, with relevance both for understanding the population structure of species within Enterobacteriaceae and for managing

AMR. The existence of farm-level differences in *E. coli* populations that persists over time, with a small number of possible interfarm transfers, suggests that livestock farms function as distinct but linked niches. It could be that “everything is everywhere” (frequent movement of strains and genes between farms), but “the environment selects” (different farms have different selective pressures). However, the observation of persistent strains over the course of a year on farms, despite presumably varying selective conditions, and the overrepresentation of putative cross-genera HGT events in isolates at the same location suggest that geographical effects or intrinsic properties of certain bacterial/MGE lineages could affect the evolution of AMR on these time scales. Future modeling work and investigation will be required to distinguish these hypotheses. Overall, our findings underline the importance of local control strategies for the emergence and spread of AMR beyond clinical settings.

Here, we have focused on genome dynamics across different niches but have not attempted to integrate our findings with



detailed contextual information from the participating sites. This will be crucial to give further insight into how differences in antimicrobial usage on farms and other management practices can affect AMR in multispecies pangenomes, such as the decline in *mcr-1* prevalence in Enterobacteriaceae on pig farms after a ban on colistin in feed, as observed in both China (40) and the UK (41). Other factors beyond antimicrobial usage could include herd size, herd management practices, and cleaning and disinfection practices. Investigating these is part of our ongoing work. Similarly, differences in management between WwTWs may affect the prevalence of AMR genes in effluent, and this is also the subject of ongoing work.

Resource limitations meant that we were unable to sequence and genetically evaluate all isolates that were cultured, and despite detailed sampling, we will not have captured all the persistence, HGT, and strain-sharing events across niches: The instances of sharing that we did identify remain fairly anecdotal. This is a problem faced by even intensive sampling efforts, which can only capture a tiny fraction of the population diversity and so are unlikely to retrieve close links in transmission chains (42) (of strains or MGEs). Future analyses could still investigate the links between such sharing and possible transmission routes between both farms and WwTWs, including the land application of manure (43), the land application of sewage sludge (44), and groundwater flow (45). Older studies have established that specific AMR genes can be more prevalent in groundwater closer to manure storage (46), but working out how to do equivalent genomic analyses for strains and MGEs across larger geographic distances with such diversity is difficult. Although this study is unprecedented in evaluating four genera in such detail, AMR gene dissemination and important structural associations of AMR genes and MGEs may also be occurring within other genera not studied here. Furthermore, we did not investigate the relationship between isolates in this study and clinical human compartments in the same study area. Ludden *et al.* (29, 47) previously reported limited direct overlap for *E. coli* or *Klebsiella pneumoniae* strains from livestock and humans in a different region of England but highlighted that 5% of human *E. coli* isolates potentially shared closely related AMR-associated MGEs with those found in livestock. We intend to see whether this finding holds in our ongoing work.

In conclusion, our study highlights the plastic and dynamic nature of AMR gene dissemination within the pangenome of major Enterobacteriaceae in several important nonclinical niches. It also demonstrates how robustly evaluating the flow of AMR genes and MGEs across highly diverse and dynamic niches is challenging even with extensive sampling. The implications of this for adequately understanding dissemination and selection of AMR genes in a “One Health” context should not be underestimated.

## MATERIALS AND METHODS

Isolates were sequenced from samples collected as part of the “The environmental RESistome: confluence of Human and Animal Biota in antibiotic resistance spread” (REHAB) study in 2017, which aimed to characterize nonclinical Enterobacteriaceae populations in four different niches within a defined study area of South Central England: cattle farms, pig farms, sheep farms, and water environments linked to WwTWs. Sampling occurred at each location at three separate TPs: January to April 2017 (TP1), June to July 2017 (TP2), and October to November 2017 (TP3).

## Farms

Five cattle farms, five sheep farms, and four pig farms were recruited from the study area following a defined recruitment process. Briefly, we aimed to recruit the five largest farms for each livestock type within the area using local APHA databases, progressively inviting the next largest farm if a farm declined. All participating farmers provided written consent for farm sampling for research purposes, and farm samples were taken between January and November 2017 on three separate visits (“TPs”) for each farm. Each farm was divided in five or fewer “epidemiological groups,” defined as a group of animals expected to share similar characteristics and managed in the same way. Ten pooled samples were collected from each of these groups, with each sample composed of small pinches of fresh feces from the floor combined into a small composite sample around 5 cm in diameter. Each group’s 10 samples were pooled, diluted up to  $10^{-5}$  in phosphate buffer solution (pH 7.2), and plated onto CHROMagar ECC (CHROMagar Microbiology, Paris, France) and CHROMagar ECC plates containing cefotaxime (1 mg/liter) as a marker for multidrug resistance. Up to 10 colonies were collected from cefotaxime (1 mg/liter)–supplemented plates and 14 colonies from CHROMagar ECC plates; where 10 colonies were not recovered, additional colonies were taken from the CHROMagar ECC plates, resulting in 24 isolates per farm. Pure isolate subcultures were subsequently stored at  $-80^{\circ}\text{C}$  in MicroBank beads (Pro-Lab Diagnostics, Neston, Cheshire, UK), and the bacterial species were confirmed using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) (Bruker, Coventry, UK) or 16S rRNA sequencing (48). The median number of sequenced isolates for a farm TP combination was 12 (range: 9 to 14), with 495 farm isolates in total: cattle ( $n = 178$ ), pig ( $n = 143$ ), and sheep ( $n = 174$ ).

## Wastewater treatment works

Five WwTWs were selected on the basis of a defined recruitment process including the following: geographic location within the study area, wastewater treatment configuration, wastewater population equivalent served, consented flow, and the accessibility of the effluent receiving river for sampling both upstream and downstream. Sampling took place in 2017 over three sampling rounds: February to March (TP1), June to July (TP2), and October to November (TP3). Sewage influent samples were collected after WwTW coarse screens, and effluent samples were collected at the last sampling point before entering the river. For each sampling round, ~6 repeated 200-ml samples of influent and effluent were collected between 9:00 a.m. and 12:00 p.m. using a sampling pole and sterile Whirl-Pak collection bags. Repeat samples in each round were pooled before processing to reduce the impact of temporal variability in wastewater flows and composition. Sediment samples were collected from 100 m upstream and 250 m downstream of the effluent entry point into the river. Sediment samples were collected using a custom sampling pole that held a removable 50-ml plastic centrifuge tube (Sigma-Aldrich, UK). Using a fresh sterile 50-ml tube each time, sediment from the riverbed was collected from the surface layer at three points at each sampling site; near bank, the center of the river, and the far bank. These samples were pooled before analysis to account for spatial variability in sediment composition. Influent, effluent, and sediment samples were stored in an insulated box at  $\sim 4^{\circ}\text{C}$  until getting back to the laboratory (<6 hours). Influent, effluent, 100 m upstream, and 250 m downstream environmental samples collected from each sewage treatment works

were transferred to the laboratory on ice and processed within 24 hours of collection. Each sample was vortexed briefly, serially diluted to  $10^{-3}$  in nutrient broth containing 10% glycerol (Oxoid, Basingstoke, UK), and plated onto CHROMagar Orientation agar (Chromagar, Paris, France) and CHROMagar Orientation agar supplemented with cefotaxime (1  $\mu\text{g}/\text{ml}$ ) (Cambridge Biosciences, Cambridge, UK). Colonies with putative morphology for species of interest were subcultured from dilution plates with suitably isolated growth. A total of up to 20 colonies was picked per sample: Up to 10 colonies were picked from the cefotaxime (1  $\text{mg}/\text{liter}$ )–supplemented plates, and the remainder were picked from the nonsupplemented plates. Pure isolates subcultured on Columbia blood agar (CBA) (Oxoid, Basingstoke, UK) were subsequently stored at  $-80^{\circ}\text{C}$  in nutrient broth containing 10% glycerol, and bacterial species were confirmed using MALDI-TOF (Bruker, Coventry, UK).

### DNA sequencing

A subset of isolates were selected for sequencing to represent diversity within the four major genera within each niche, including the use of third-generation cephalosporin resistance as a selective marker to identify a subgroup of multidrug-resistant isolates within each genus. A total of 832 isolates were each sequenced with both a short-read (Illumina HiSeq 4000) and a long-read sequencing approach [four isolates selected for sequencing failed subsequent hybrid assembly and were not included in further analyses; one further isolate was removed after assembly (see “Genome assembly”)]. For the first TP, the latter involved sequencing using either PacBio SMRT ( $n = 192$ ) or Oxford Nanopore Technologies (ONT) methodologies ( $n = 127$ ). The results of a pilot study comparing sequencing and assembly approaches using a subset of REHAB isolates (28) were used to inform the choice of ONT as the long-read sequencing approach for all isolates from the second ( $n = 255$ ) and third ( $n = 254$ ) TPs.

Isolate stocks from  $-80^{\circ}\text{C}$  storage were cultured onto CBA and supplemented with cefpodoxime (Thermo Fisher Scientific, USA) 10- $\mu\text{g}$  discs for isolates not sensitive to cefotaxime during original sample isolation. DNA was extracted using the Qiagen Genomic tip/100G (Qiagen, Venlo, The Netherlands) according to the manufacturer’s instructions. DNA concentration was quantified by Qubit 2.0 fluorimeter (Invitrogen, UK), and quality and fragment size distribution were assessed by TapeStation 2200 (Agilent, Santa Clara, USA). ONT sequencing libraries were prepared by multiplexing six to eight DNA extracts per flow cell using kits SQK-RBK004, SQK-LSK108, and EXP-NBD103 according to the manufacturer’s protocol. Libraries were loaded onto flow cell versions FLO-MIN106 R9.4 (0.1) SpotON and sequenced for 48 hours on a GridION (ONT, Oxford, UK).

### Genome assembly

We used the hybrid assembly and sequencing methods described in our pilot study (28) to produce high-quality Enterobacteriaceae genomes from short and long reads. Briefly, we used Unicycler (v0.4.7) (49) with “normal” mode, `--min_component_size 500`, `--min_dead_end_size 500`, and otherwise default parameters. Our pilot study (28) explored the accuracy of assemblies from this method (and others) in detail using multiple metrics, including the following: the mapping of Illumina short reads back to assemblies, the mapping of long reads back to assemblies, the comparison of assemblies from the same isolate, using known marker genes to assess overall quality, the assembly’s overall “circularity,” and the presence

of indel errors. We found excellent concordance in structural agreement of circular structures from different assemblies [see, e.g., fig. S4 of (28)], giving us confidence that circular structures from hybrid assemblies (chromosomes and plasmids) are accurate. One possible concern could be that chimeric reads in long-read datasets could lead to incorrect structures in the final hybrid assembly. However, an independent investigation (50, 51) using data from our pilot study alongside simulated data showed that long read-only assemblies with Unicycler are only affected by the presence of chimeric reads at rates of  $\sim 15\%$  of reads and are thus not a concern for real datasets, which typically have rates of  $<2\%$ .

Final assemblies from all isolates had a median of four contigs [interquartile range (IQR): 3 to 8; range: 1 to 391], with a median of two circularised plasmids (IQR: 1 to 4, range: 0 to 14). One *Citrobacter* isolate from TP1 was removed from the dataset after we identified a sample mixup, meaning that its metadata were unreliable. The majority (616 of 827, 74.5%) of the assemblies had a circularized chromosome, and 558 of 827 (67.4%) were complete, i.e., chromosome and all plasmids circularized (table S1).

### Genome assignment and typing

We assigned species and ST from assembled genomes using mlst (v2.16.4) (52). We also validated species assignments by downloading all National Center for Biotechnology Information (NCBI) Refseq complete genomes for the four genera under study as of 4 June 2020 and using fastANI (v1.3) (53) to compute average nucleotide identity scores against reference genomes for each assembled genome. We took the species assignment of the top hit for each assembled genome. Furthermore, we manually checked genus assignments using a t-distributed stochastic neighbor embedding (t-SNE) plot of isolate genomes against a collection of reference genomes and made corrections to the assignment if necessary. We used ClermonTyping (v1.4.1) (54) to assign phylogroup to  $n = 553$  *Escherichia* isolates. Considering the genus *Escherichia*, there were 553 isolates, 410 with circularized chromosomes, and of these, 379 were complete genomes containing 961 complete plasmids in total. Considering only *E. coli*, there were 502 *E. coli* isolates, 372 with circularized chromosomes, and of these, 348 were complete genomes containing 878 complete plasmids in total. A minority of genomes were *E. fergusonii* ( $n = 51$ ), from clades I to V ( $n = 14$ ), or could not be typed ( $n = 7$ ), with  $n = 481$  genomes from within the principal *E. coli* phylogroups (A:  $n = 131$ ; B1:  $n = 193$ ; B2:  $n = 59$ ; C:  $n = 11$ ; D:  $n = 25$ ; E:  $n = 50$ ; F:  $n = 6$ ; and G:  $n = 6$ ).

Sequenced isolates from three other Enterobacteriaceae genera included the following: *Citrobacter* ( $n = 128$ : 81 *Citrobacter freundii* and 46 unassigned *Citrobacter* sp.), *Enterobacter* ( $n = 71$ : 59 *Enterobacter cloacae* and 12 unassigned *Enterobacter* sp.), and *Klebsiella* ( $n = 76$ : 40 *K. pneumoniae*, 30 *Klebsiella oxytoca*, 2 *Klebsiella aerogenes*, and 4 unassigned *Klebsiella* sp.). The majority of the farm-associated isolates were *E. coli*, whereas WwTW-associated isolates had roughly equal numbers of genera (fig. S1). This reflects both the diversity present in each niche and the selection strategy to sequence equal numbers across genera where feasible.

### Pangenome analysis

All genomes were annotated with Prokka (v1.14.0) (55). We performed a multispecies pangenome analysis by clustering genes into gene groups using Roary (v3.12.0) (56) across all isolates at various sequence identity thresholds with the maximum number of clusters

set to 300,000 (-g 300,000) and without splitting paralogs (-s). At a 95% identity for blastp, there were 139,788 gene groups across all genera. Further to this analysis, genes were also clustered at a higher sequence identity (>99% identity threshold) to identify recent HGT events, which gave 214,743 gene groups across all genera. (These pangenome analyses included the subsequently removed *Citrobacter* isolate with unreliable metadata.) For  $n = 616$  isolates with circularised chromosomes, we split the genome into chromosomal and plasmid-borne components (i.e., all other contigs) to analyze the genomic location of genes. We excluded isolates without circularized chromosomes from this analysis. For within-species pangenome analyses, the more recently developed Panaroo gives lower annotation error rates and a more accurate core genome than Roary or other methods by using gene adjacency (i.e., synteny) information (57), although it is not suitable for cross-species analyses where no core genome is expected. Therefore, for a higher-resolution within-species analysis of  $n = 488$  *E. coli* isolates (excluding *E. fergusonii* and clades I to V), we used Panaroo (v0.1.0) (57) to extract a core genome alignment based on 2915 concatenated core genes (Fig. 1C). The phylogeny was produced using iqtree (v1.6.11) (58), with branch lengths not corrected for recombination, and plotted with ggtree (v2.0.1) (59).

### Plasmid annotation and clustering

We searched all plasmids against PLSDb (version: 2020-03-04) (60), which contains 20,668 complete published plasmids, using “screen” in mash (v2.0) (61) and keeping the top hit. All plasmids had a match apart from 17 small plasmids predicted to be non-mobilizable (median length: 4.8 kbp; range: 2.9 to 20.7 kbp), from *Escherichia* ( $n = 11$ ), *Enterobacter* ( $n = 2$ ), and *Citrobacter* ( $n = 4$ ). We clustered plasmids using mob cluster and assigned replicon types with mob typer, both part of the MOB suite (v1.4.9) (62). Mob cluster uses single linkage clustering with a cutoff of a mash distance of 0.05 [corresponding to 95% average nucleotide identity (ANI)], resulting in 611 clusters (table S2). In total, there were 134 different combinations of replicons observed on plasmids (“replicon haplotypes”). The most abundant replicon was IncFIB ( $n = 459$ ), which was seen across all niches [pig ( $n = 80$ ), cattle ( $n = 113$ ), sheep ( $n = 78$ ), and WWTWs ( $n = 188$ )]. Only nine small multicopy plasmids (~6 kbp) carried AMR genes, all of which had a ColRNAI replicon; these ColRNAI plasmids have been proposed to be sources of evolutionary innovation (37, 38).

We considered the relationship between such “distance-free” clustering and plasmid gene content. On the basis of gene clustering with Roary (see above), we compared the structure of circularised plasmids using all connecting edges between two genes. We defined the resemblance for both gene content (gene presence/absence) and gene structure. The gene content resemblance between two plasmids with  $n_1$  and  $n_2$  genes, respectively, with  $N$  genes in common, was defined as  $r_{\text{content}} = 2N/(n_1 + n_2)$ . The edge structure resemblance between two plasmids with  $g$  gene-gene edges in common was defined as  $r_{\text{edge}} = 2g/(n_1 + n_2)$ . Typically  $r_{\text{edge}} < r_{\text{content}}$  but this definition does allow for the case where repeated genetic elements produce  $r_{\text{edge}} > r_{\text{content}}$  (e.g., fig. S6B).

### Comparison of plasmid-borne and chromosomal pangenome components

To visualize cross-genera pangenomes (e.g., Fig. 2), we used t-SNE. We used the Rtsne function with a perplexity of 30 on gene presence/absence matrices in the Rtsne R package (v0.15) (63). To

conduct permutational analyses of variance, we used the adonis function from the vegan R package (v2.5-6) (64) on the matrix of pairwise Jaccard distances, which was calculated using the vegdist function. For between-genera analyses, we used the formula  $\text{dist} \sim \text{niche} * \text{genus}$ . For within-*Escherichia* analyses, we used the formula  $\text{dist} \sim \text{niche} * \text{phylogroup}$ .

### Detection of AMR genes and ISs

We searched assemblies using ABRicate (v0.9.8) (65) for acquired resistance genes (i.e., excluding mutational resistance) in the NCBI AMRFinder Plus database (PRJNA313047). We used a minimum identity threshold of 90% and a minimum coverage threshold of 90% (table S4). Isolates that cultured selectively from cefotaxime-supplemented plates carried more AMR genes than nonselectively cultured isolates (median of 7.5 versus 1.0), as expected. We also searched for ISs using the ISFinder database (66) as a database in ABRicate with the same identity and coverage thresholds (table S5).

### Detection of recent HGT events

We performed an all-against-all comparison of assemblies with mummer (v3.23-2) (67) using the -maxmatch option to identify shared sequences of length >5000 bp between genomes of different genera (these could include both transfer of whole plasmids or partial sequences). For comparing the observed distribution of cross-genera HGT events to the expected, we assumed a random distribution drawn from all possible cross-genera comparisons from livestock isolates.

### Distribution of ISs

We constructed the bipartite presence/absence network of ISs and replicon haplotypes for the 34 replicon haplotypes, which were observed on 10 or more plasmids. We simulated null models of co-occurrence patterns using the cooc\_null\_model with null model sim9, which fixes the row and column sums of the presence/absence matrix, in the R package EcoSimR (v0.1.0) (68). Simulations used  $n = 10,000$  iterations with a burn-in of 500 iterations.

### Modeling of GRR

We selected a subset of *E. coli* genomes with a circularized chromosome ( $n = 363$ ) and used the core genome tree constructed with iqtree (Fig. 1; dropping other *E. coli* isolates) to calculate patristic distances between isolates. We calculated chromosomal and plasmid GRR for all pairwise comparisons using output from roary (95% identity threshold, as above) and fit linear models for GRR (Fig. 4).

### SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/7/15/eabe3868/DC1>

### REFERENCES AND NOTES

1. N. Stoesser, A. E. Sheppard, C. E. Moore, T. Golubchik, C. M. Parry, P. Nget, M. Saroeun, N. P. Day, A. Giess, J. R. Johnson, T. E. Peto, D. W. Crook, A. S. Walker; Modernizing Medical Microbiology Informatics Group, Extensive within-host diversity in fecally carried extended-spectrum-beta-lactamase-producing *Escherichia coli* isolates: Implications for transmission analyses. *J. Clin. Microbiol.* **53**, 2122–2131 (2015).
2. L. Poirel, J.-Y. Madec, A. Lupo, A.-K. Schink, N. Kieffer, P. Nordmann, S. Schwarz, Antimicrobial resistance in *Escherichia coli*. *Microbiol. Spectr.* **6**, (2018).
3. J. Jang, H. G. Hur, M. J. Sadowsky, M. N. Byappanahalli, T. Yan, S. Ishii, Environmental *Escherichia coli*: Ecology and public health implications—a review. *J. Appl. Microbiol.* **123**, 570–581 (2017).



4. N. Mahfouz, S. Caucci, E. Achatz, T. Semmler, S. Guenther, T. U. Berendonk, M. Schroeder, High genomic diversity of multi-drug resistant wastewater *Escherichia coli*. *Sci. Rep.* **8**, 8928 (2018).
5. A. Leimbach, J. Hacker, U. Dobrindt, *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* **358**, 3–32 (2013).
6. L. K. Logan, R. A. Weinstein, The epidemiology of carbapenem-resistant Enterobacteriaceae: The impact and evolution of a global menace. *J. Infect. Dis.* **215**, S28–S36 (2017).
7. J. Iredell, J. Brown, K. Tagg, Antibiotic resistance in Enterobacteriaceae: Mechanisms and clinical implications. *BMJ* **352**, h6420 (2016).
8. A. E. Sheppard, N. Stoesser, D. J. Wilson, R. Sebra, A. Kasarskis, L. W. Anson, A. Giess, L. J. Pankhurst, A. Vaughan, C. J. Grim, H. L. Cox, A. J. Yeh; the Modernising Medical Microbiology (MMM) Informatics Group, C. D. Sifri, A. S. Walker, T. E. Peto, D. W. Crook, A. J. Mathers, Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrob. Agents Chemother.* **60**, 3767–3778 (2016).
9. R. Wang, L. van Dorp, L. P. Shaw, P. Bradley, Q. Wang, X. Wang, L. Jin, Q. Zhang, Y. Liu, A. Rieux, T. Dorai-Schneiders, L. A. Weinert, Z. Iqbal, X. Didelot, H. Wang, F. Balloux, The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* **9**, 1179 (2018).
10. D. Skurnik, R. Ruimy, A. Andremont, C. Amorin, P. Rouquet, B. Picard, E. Denamur, Effect of human vicinity on antimicrobial resistance and integrons in animal faecal *Escherichia coli*. *J. Antimicrob. Chemother.* **57**, 1215–1219 (2006).
11. C. F. Nnadozie, O. N. Odume, Freshwater environments as reservoirs of antibiotic resistant bacteria and their role in the dissemination of antibiotic resistance genes. *Environ. Pollut. Barking Essex* **254**, 113067 (2019).
12. M. Woolhouse, M. Ward, B. van Bunnik, J. Farrar, Antimicrobial resistance in humans, livestock and the wider environment. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140083 (2015).
13. M. AbuOun, H. M. O'Connor, E. J. Stubberfield, J. Nunez-Garcia, E. Sayers, D. W. Crook, R. P. Smith, M. F. Anjum, Characterizing antimicrobial resistant *Escherichia coli* and associated risk factors in a cross-sectional study of pig farms in Great Britain. *Front. Microbiol.* **11**, 861 (2020).
14. K. L. Wyres, M. M. C. Lam, K. E. Holt, Population genomics of *Klebsiella pneumoniae*. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).
15. D. A. Rasko, M. J. Rosovitz, G. S. A. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebailhia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, J. Ravel, The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893 (2008).
16. E. N. Gordienko, M. D. Kazanov, M. S. Gelfand, Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*. *J. Bacteriol.* **195**, 2786–2792 (2013).
17. M. Juhás, Horizontal gene transfer in human pathogens. *Crit. Rev. Microbiol.* **41**, 101–108 (2013).
18. M. A. Brockhurst, E. Harrison, J. P. J. Hall, T. Richards, A. McNally, C. MacLean, The ecology and evolution of pangenomes. *Curr. Biol.* **29**, R1094–R1103 (2019).
19. B. J. Shapiro, The population genetics of pangenomes. *Nat. Microbiol.* **2**, 1574 (2017).
20. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
21. I. Sela, Y. I. Wolf, E. V. Koonin, Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11399–11407 (2016).
22. N. A. Andreani, E. Hesse, M. Vos, Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
23. A. G. Decano, T. Downing, An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.* **9**, 17394 (2019).
24. N. K. Petty, N. L. Ben Zakour, M. Stanton-Cook, E. Skippington, M. Totsika, B. M. Forde, M. D. Phan, D. Gomes Moriel, K. M. Peters, M. Davies, B. A. Rogers, G. Dougan, J. Rodriguez-Bano, A. Pascual, J. D. D. Pitout, M. Upton, D. L. Paterson, T. R. Walsh, M. A. Schembri, S. A. Beatson, Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5694–5699 (2014).
25. O. Tenaillon, D. Skurnik, B. Picard, E. Denamur, The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* **8**, 207–217 (2010).
26. R. C. MacLean, A. San Millan, Microbial evolution: Towards resolving the plasmid paradox. *Curr. Biol. CB* **25**, R764–R767 (2015).
27. S. George, L. Pankhurst, A. Hubbard, A. Votintseva, N. Stoesser, A. E. Sheppard, A. Mathers, R. Norris, I. Navickaite, C. Eaton, Z. Iqbal, D. W. Crook, H. T. T. Phan, Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: Assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microb. Genomics* **3**, e000118 (2017).
28. N. De Maio, L. P. Shaw, A. Hubbard, S. George, N. D. Sanderson, J. Swann, R. Wick, M. A. Oun, E. Stubberfield, S. J. Hoosdally, D. W. Crook, T. E. A. Peto, A. E. Sheppard, M. J. Bailey, D. S. Read, M. F. Anjum, A. S. Walker, N. Stoesser; Rehab Consortium, Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* **5**, e000294 (2019).
29. C. Ludden, K. E. Raven, D. Jamroz, T. Gouliouris, B. Blane, F. Coll, M. de Goffau, P. Naydenova, C. Horner, J. Hernandez-Garcia, P. Wood, N. Hadjirin, M. Radakovic, N. M. Brown, M. Holmes, J. Parkhill, S. J. Peacock, One health genomic surveillance of *Escherichia coli* demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. *MBio* **10**, e02693-18 (2019).
30. M. Touchon, A. Perrin, J. A. M. de Sousa, B. Vangchhia, S. Burn, C. L. O'Brien, E. Denamur, D. Gordon, E. P. C. Rocha, Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *bioRxiv*, 2020.02.12.945709 (2020).
31. S. Redondo-Salvo, R. Fernández-López, R. Ruiz, L. Vielva, M. de Toro, E. P. C. Rocha, M. P. Garcillán-Barcia, F. de la Cruz, Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
32. A. San Millan, K. Heilbron, R. C. MacLean, Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. *ISME J.* **8**, 601–612 (2014).
33. S. He, A. B. Hickman, A. M. Varani, P. Siguier, M. Chandler, J. P. Dekker, F. Dyda, Insertion sequence is26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio* **6**, e00762 (2015).
34. K. H. Hansen, M. R. Andreassen, M. S. Pedersen, H. Westh, L. Jelsbak, K. Schønning, Resistance to piperacillin/tazobactam in *Escherichia coli* resulting from extensive IS26-associated gene amplification of blaTEM-1. *J. Antimicrob. Chemother.* **74**, 3179–3183 (2019).
35. D. D. He, S. Y. Zhao, H. Wu, G. Z. Hu, J. F. Zhao, Z. Y. Zong, Y. S. Pan, Antimicrobial resistance-encoding plasmid clusters with heterogeneous MDR regions driven by IS26 in a single *Escherichia coli* isolate. *J. Antimicrob. Chemother.* **74**, 1511–1516 (2019).
36. S. Zhi, G. Banting, Q. Li, T. A. Edge, E. Topp, M. Sokurenko, C. Scott, S. Braithwaite, N. J. Ruecker, Y. Yasui, T. McAllister, L. Chui, N. F. Neumann, Evidence of Naturalized Stress-Tolerant Strains of *Escherichia coli* in Municipal Wastewater Treatment Plants. *Appl. Environ. Microbiol.* **82**, 5505–5518 (2016).
37. A. San Millan, J. A. Escudero, D. R. Gifford, D. Mazer, R. C. MacLean, Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.* **1**, 0010 (2017).
38. J. Rodriguez-Beltran, J. C. R. Hernandez-Beltran, J. DelaFuente, J. A. Escudero, A. Fuentes-Hernandez, R. C. MacLean, R. Peña-Miller, A. San Millan, Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation. *Nat. Ecol. Evol.* **2**, 873–881 (2018).
39. A.-K. Dietel, H. Merker, M. Kaltenpoth, C. Kost, Selective advantages favour high genomic AT-contents in intracellular elements. *PLOS Genet.* **15**, e1007778 (2019).
40. C. Shen, L. Z. Zhong, Y. Yang, Y. Doi, D. L. Paterson, N. Stoesser, F. Ma, M. A. E. G. el-Sayed Ahmed, S. Feng, S. Huang, H. Y. Li, X. Huang, X. Wen, Z. Zhao, M. Lin, G. Chen, W. Liang, Y. Liang, Y. Xia, M. Dai, D. Q. Chen, L. Zhang, K. Liao, G. B. Tian, Dynamics of mcr-1 prevalence and mcr-1-positive *Escherichia coli* after the cessation of colistin use as a feed additive for animals in China: A prospective cross-sectional and whole genome sequencing-based molecular epidemiological study. *Lancet Microbe* **1**, e34–e43 (2020).
41. N. A. Duggett, L. P. Randall, R. A. Horton, F. Lemma, M. Kirchner, J. Nunez-Garcia, C. Brena, S. M. Williamson, C. Teale, M. F. Anjum, Molecular epidemiology of isolates with multiple mcr plasmids from a pig farm in Great Britain: The effects of colistin withdrawal in the short and long term. *J. Antimicrob. Chemother.* **73**, 3025–3033 (2018).
42. W. P. Hanage, Two health or not two health? That is the question. *MBio* **10**, e00550-19 (2019).
43. J. C. Chee-Sanford, R. I. Mackie, S. Koike, I. G. Krapac, Y. F. Lin, A. C. Yannarell, S. Maxwell, R. I. Aminov, Fate and transport of antibiotic residues and antibiotic resistance genes following land application of manure waste. *J. Environ. Qual.* **38**, 1086–1108 (2009).
44. Q. Chen, X. An, H. Li, J. Su, Y. Ma, Y.-G. Zhu, Long-term field application of sewage sludge increases the abundance of antibiotic resistance genes in soil. *Environ. Int.* **92–93**, 1–10 (2016).
45. L. Andrade, M. Kelly, P. Hynds, J. Weatherill, A. Majury, J. O'Dwyer, Groundwater resources as a global reservoir for antimicrobial-resistant bacteria. *Water Res.* **170**, 115360 (2020).
46. R. I. Mackie, S. Koike, I. Krapac, J. Chee-Sanford, S. Maxwell, R. I. Aminov, Tetracycline residues and tetracycline resistance genes in groundwater impacted by swine production facilities. *Anim. Biotechnol.* **17**, 157–176 (2006).
47. C. Ludden, D. Moradigaravand, D. Jamroz, T. Gouliouris, B. Blane, P. Naydenova, J. Hernandez-Garcia, P. Wood, N. Hadjirin, M. Radakovic, C. Crawley, N. M. Brown, M. Holmes, J. Parkhill, S. J. Peacock, A One Health study of the genetic relatedness of *Klebsiella pneumoniae* and their mobile elements in the East of England. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **70**, 219–226 (2020).
48. K. J. Edwards, J. M. J. Logan, S. Langham, C. Swift, S. E. Gharbia, Utility of real-time amplification of selected 16S rRNA gene sequences as a tool for detection and identification of microbial signatures directly from clinical samples. *J. Med. Microbiol.* **61**, 645–652 (2012).
49. R. R. Wick, L. M. Judd, C. L. Gorrie, K. E. Holt, Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
50. R. R. Wick, K. E. Holt, Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Res.* **8**, 2138 (2019).
51. Ryan Wick, *rrwick/Long-read-assembler-comparison: Initial release*. (Zenodo, 2019). doi:10.5281/zenodo.2702443.

52. T. Seemann, *mlst*; <https://github.com/tseemann/mlst>.
53. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
54. J. Beghain, A. Bridier-Nahmias, H. Le Nagard, E. Denamur, O. Clermont, ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb. Genomics* **4**, e000192 (2018).
55. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
56. A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, J. Parkhill, Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
57. G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. W. Frost, J. Corander, S. D. Bentley, J. Parkhill, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
58. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
59. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
60. V. Galata, T. Fehlmann, C. Backes, A. Keller, PLSDb: A resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).
61. B. D. Ondov, G. J. Starrett, A. Sappington, A. Kostic, S. Koren, C. B. Buck, A. M. Phillippy, Mash Screen: High-throughput sequence containment estimation for genome discovery. *bioRxiv*, 557314 (2019).
62. J. Robertson, J. H. E. Nash, MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genomics* **4**, e000206 (2018).
63. J. H. Krijthe, *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation* (2015).
64. J. Oksanen et al. *vegan: Community Ecology Package* (2019).
65. T. Seemann, *Abricate* (2020).
66. P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, M. Chandler, ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
67. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
68. N. J. Gotelli, E. M. Hart, A. M. Ellison, *EcoSimR: Null Model Analysis for Ecological Data. R Package Version 0.1.0.* (2015).
69. K. A. Jolley, M. C. Maiden, BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).

## Acknowledgments

**Funding:** This work was funded by the AMR Cross-council Initiative supported by the seven research councils [grants NE/N019989/1 and NE/N019660/1]. Work was also supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in

Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford (NIHR200915) in partnership with Public Health England (PHE), and by the NIHR Oxford Biomedical Research Centre (BRC). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from a Wellcome Trust Core Award Grant (grant 203141/Z/16/Z). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. K.K.C. is Medical Research Foundation-funded. D.W.C., T.E.A.P., and A.S.W. are NIHR senior investigators. Image credits: The following images are used in Figure 1: petri dish icon made by monkik; pig, cow, sheep, and feces icons made by Freepik; and WwTW symbol made by Smashicons (all sourced from flaticon.com). PubMLST: This publication made use of the PubMLST website (<https://pubmlst.org/>) developed by K. Jolley(69) and sited at the University of Oxford. The development of that website was funded by the Wellcome Trust. **Author contributions:** Using the CRediT system, author contributions were as follows: conceptualization: L.P.S., M.J.Ba., D.W.C., D.S.R., M.F.A., A.S.W., and N.S.; methodology: L.P.S., A.S.W., and N.S.; software: L.P.S. and J.S.; validation: L.P.S., K.K.C., J.K., M.A., E.S., L.B., G.R., A.T.M.H., H.P., R.S., and N.S.; formal analysis: L.P.S.; investigation: K.K.C., J.K., M.A., E.S., H.S.G., L.B., G.R., M.J.Bo., A.T.M.H., H.P., J.S., D.G., R.P.S., R.S., and N.S.; resources: all authors; data curation: L.P.S., K.K.C., J.K., M.A., E.S., L.B., G.R., A.T.M.H., H.P., J.S., D.G., R.P.S., R.S., and N.S.; writing (original draft): L.P.S., A.S.W., and N.S.; writing (review and editing): all authors; visualization: L.P.S.; supervision: T.E.A.P., M.J.Ba., D.W.C., D.S.R., M.F.A., A.S.W., and N.S.; project administration: M.A., R.P.S., S.J.H., D.S.R., M.F.A., A.S.W., and N.S.; and funding acquisition: M.J.B., D.W.C., D.S.R., M.F.A., and N.S. The REHAB consortium is represented by the following: M. Abuoun, M. F. Anjum, M. J. Bailey, H. Brett, M. Bowes, K. K. Chau, D. W. Crook, N. de Maio, D. Gilson, S. George, H. S. Gweon, A. Hubbard, S. J. Hoosdally, J. Kavanagh, H. Jones, W. Matlock, T. E. A. Peto, D. S. Read, R. Sebra, L. P. Shaw, A. E. Sheppard, R. Smith, E. Stubberfield, N. Stoesser, J. Swann, A. S. Walker, and N. Woodford. **Competing interests:** R.S. is the VP of Technology Development at Sema4, a Mount Sinai venture. Other authors declare that they have no competing interests. **Data and materials availability:** Sequencing data and assemblies for all isolates are available on NCBI under BioProject accession PRJNA605147. BioSample, Sequencing Read Archive (SRA), and Assembly accessions for all isolates are provided in table S1. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 18 August 2020

Accepted 22 February 2021

Published 9 April 2021

10.1126/sciadv.abe3868

**Citation:** L. P. Shaw, K. K. Chau, J. Kavanagh, M. Abuoun, E. Stubberfield, H. S. Gweon, L. Barker, G. Rodger, M. J. Bowes, A. T. M. Hubbard, H. Pickford, J. Swann, D. Gilson, R. P. Smith, S. J. Hoosdally, R. Sebra, H. Brett, T. E. A. Peto, M. J. Bailey, D. W. Crook, D. S. Read, M. F. Anjum, A. S. Walker, N. Stoesser, REHAB consortium, Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Sci. Adv.* **7**, eabe3868 (2021).



## Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae

Liam P. Shaw, Kevin K. Chau, James Kavanagh, Manal AbuOun, Emma Stubberfield, H. Soon Gweon, Leanne Barker, Gillian Rodger, Mike J. Bowes, Alasdair T. M. Hubbard, Hayleah Pickford, Jeremy Swann, Daniel Gilson, Richard P. Smith, Sarah J. Hoosdally, Robert Sebra, Howard Brett, Tim E. A. Peto, Mark J. Bailey, Derrick W. Crook, Daniel S. Read, Muna F. Anjum, A. Sarah Walker, Nicole Stoesser and REHAB consortium

*Sci Adv* 7 (15), eabe3868.  
DOI: 10.1126/sciadv.abe3868

### ARTICLE TOOLS

<http://advances.sciencemag.org/content/7/15/eabe3868>

### SUPPLEMENTARY MATERIALS

<http://advances.sciencemag.org/content/suppl/2021/04/05/7.15.eabe3868.DC1>

### REFERENCES

This article cites 60 articles, 10 of which you can access for free  
<http://advances.sciencemag.org/content/7/15/eabe3868#BIBL>

### PERMISSIONS

<http://www.sciencemag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

*Science Advances* (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).